

(19) World Intellectual Property Organization  
International Bureau



(43) International Publication Date  
27 February 2003 (27.02.2003)

PCT

(10) International Publication Number  
**WO 03/015613 A2**

(51) International Patent Classification<sup>7</sup>: **A61B**

(21) International Application Number: PCT/US02/26027

(22) International Filing Date: 16 August 2002 (16.08.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:  
60/312,400 16 August 2001 (16.08.2001) US

(71) Applicants (*for all designated States except US*): **THE UNITED STATES OF AMERICA AS REPRESENTED BY THE SECRETARY OF HEALTH AND HUMAN SERVICES** [US/US]; NIH Office of Technology Transfer, 6011 Executive Boulevard, Suite 325, Rockville, MD 20852 (US). **GENZYME CORPORATION** [US/US]; One Kendall Square, Cambridge, MA 02139 (US). **THE JOHN HOPKINS UNIVERSITY SCHOOL OF MEDICINE** [US/US]; 111 Market Place, Suite 906, Baltimore, MD 21202 (US).

(72) Inventors; and

(75) Inventors/Applicants (*for US only*): **NACHT, Mariana** [US/US]; c/o Genzyme Corporation, One Kendall Square, Cambridge, MA 02139 (US). **DRACHEVA, Tatiana** [US/US]; c/o The United States of America as represented by, the Secretary of Health and Human Service, NIH Office of Technology Transfer, 6011 Executive Boulevard, Suite 325, Rockville, MD 20852 (US). **SIDRANSKY, David** [US/US]; c/o The John Hopkins University School of Medicine, 111 Market Place, Suite 906, Baltimore,

MD 21202 (US). **MADDEN, Stephen, L.** [US/US]; c/o Genzyme Corporation, One Kendall Square, Cambridge, MA 02139 (US). **JEN, Jin** [US/US]; The United States of America as represented by the, Secretary of Health and Human Services, NIH Office of Technology Transfer, 6011 Executive Boulevard, Suite 325, Rockville, MD 20852 (US).

(74) Agent: **KAGAN, Sarah, A.**; Banner & Witcoff, Ltd., 1001 G. Street, N.W., Suite 1100, Washington, DC 20001 (US).

(81) Designated States (*national*): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States (*regional*): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

**Published:**

— *without international search report and to be republished upon receipt of that report*

*For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.*

(54) Title: MOLECULAR CHARACTERISTICS OF NON-SMALL CELL LUNG CANCER

(57) Abstract: We used hierarchical clustering to examine gene expression profiles generated by serial analysis of gene expression (SAGE) in a total of nine normal lung epithelial cells and non-small cell lung cancers (NSCLC). Separation of normal and tumor samples, as well as histopathological subtypes, was evident using the 3,921 most abundant transcript tags. This distinction remained when just 115 highly differentially expressed transcript tags were used. Furthermore, these 115 transcript tags clustered into groups that were suggestive of the unique biological and pathological features of the different tissues examined. Adenocarcinomas were characterized by high-level expression of small airway-associated or immunologically related proteins, while squamous cell carcinomas overexpressed genes involved in cellular detoxification or antioxidation. The messages of two p53-regulated genes, p21<sup>WAF1</sup> and 14-3-3 $\sigma$ , were consistently under-expressed in the adenocarcinomas, suggesting that the p53 pathway itself might be compromised in this cancer type. Gene expression observed by SAGE were consistent with the results obtained by quantitative real-time PCR or cDNA array analyses using 43 additional lung tumor and normal samples. Thus, although derived from only a few tissue libraries, molecular signatures of non-small cell lung cancer derived from SAGE most likely represent an unbiased yet distinctive molecular signature for human lung cancer.



WO 03/015613 A2

**MOLECULAR CHARACTERISTICS OF NON-SMALL CELL LUNG CANCER**

- [01] This application claims priority to provisional U.S. Application Ser. No. 60/312,400, filed August 16, 2001.

**FIELD OF THE INVENTION**

- [02] The invention relates to the field of cancer. In particular it relates to the areas of diagnostics and lung cancer.

**BACKGROUND OF THE INVENTION**

- [03] Lung cancer is the leading cause of cancer death worldwide and NSCLC accounts for nearly 80% of the disease (1). Based on cell morphology, adenocarcinoma and squamous are the most common types of NSCLC (2). Although the clinical courses of these tumors are similar, adenocarcinomas are characterized by peripheral location in the lung and often have activating mutations in the K-ras oncogene (3, 4). In contrast, squamous cell carcinomas are usually centrally located and more frequently carry p53 gene mutations (5). Furthermore, the etiology of squamous cell carcinoma is closely associated with tobacco smoking while the cause of adenocarcinoma remains unclear (6, 7). Although many molecular changes associated with NSCLC have been reported (8, 9), the global gene expression pattern associated with these two most common types of lung cancer has not been described. Understanding gene expression patterns in these major tumor types will uncover novel markers for disease detection as well as potential targets for rational therapy of lung cancer.
- [04] Several technologies are currently being utilized for gene expression profiling in human cancer (10). SAGE (11) is an open system that rapidly identifies any expressed transcript in a tissue of interest, including transcripts that had not been identified. This highly quantitative method can accurately identify the degree of expression for each transcript.

Comparing SAGE profiles between the tumor and the corresponding normal tissues can readily identify genes differentially expressed in the two populations. Using this method, novel transcripts and molecular pathways have been discovered (12-14). In contrast, cDNA arrays represent a closed system that analyze relative expression levels of previously known genes or transcripts (15, 16). Because many thousands of genes can be placed on a single membrane or slide for rapid screening, studies have recently demonstrated molecular profiles of several human cancers (17-20).

- [05] Hierarchical clustering is a systematic method widely used in cDNA array data analysis where the difference between the expression patterns of many genes is generally within a few fold (21). We reasoned that because SAGE is highly quantitative, hierarchical clustering might be used to organize gene expression data generated by SAGE from just a few tissue libraries. To test this, SAGE tags from two of each libraries derived from primary adenocarcinomas, primary squamous cell carcinomas, normal lung small airway epithelial cells (SAEC), or normal bronchial/tracheal epithelial (NHBE) cells, and a lung adenocarcinoma cell line were used. SAGE tags showing the highest abundance were subjected to clustering analysis. Although each library was derived from a different individual, normal and tumor samples clustered in two separate branches while tissues of different cell types clustered together. Furthermore, SAGE tags clustered into biologically meaningful groups revealing the important molecular characteristics of these two most common NSCLC subtypes.

#### **BRIEF SUMMARY OF THE INVENTION**

- [06] The present invention provides a method of identifying a lung cancer as squamous cell carcinoma. According to the method an amount of a gene product of a gene in a lung cancer sample is determined. The gene is selected from the group consisting of : glutathione peroxidase (GPX; NM\_002083), glutathione S-transferase M3 (GSTM3; NM\_000849), aldoketoreductase family 1, member B 10 (NM\_020299), peroxiredoxin 1 (PRDX1; NM\_002574), small proline-rich protein 3 (SPRR3; NM\_005416), and TNF receptor superfamily member 18 (TNFRSF18; NM004195). The amount of the gene product in the lung cancer sample is compared to the amount determined in a lung tissue sample which is non-pathological. An increased amount of the gene product in the lung

cancer sample relative to the lung tissue sample which is non-pathological identifies the lung cancer as a squamous cell carcinoma.

- [07] The present invention provides a method of identifying a lung cancer as adenocarcinoma. According to the method an amount of a gene product of a small proline-rich protein 3 (SPRR3; NM\_005416) gene in a lung cancer sample is determined. The amount of the gene product in the lung cancer sample is compared to the amount determined in a lung tissue sample which is non-pathological. A decreased amount of the gene product in the lung cancer sample relative to the lung tissue sample which is non-pathological identifies the lung cancer as adenocarcinoma.
- [08] The invention thus provides the art with a molecular diagnostic to supplement or replace histological features and/or clinical behavior.

#### **BRIEF DESCRIPTION OF THE DRAWINGS**

- [09] **Fig. 1A- Fig. 1C show clustering and multidimensional scaling of the SAGE libraries.** Only genes with total tag-counts of at least 10 are included. (Fig. 1 A) Cluster of all nine SAGE libraries. Genes are aligned horizontally, libraries are shown vertically. Red, green and black colors indicate genes expressed at high, low, or moderate levels, respectively, in the indicated library. (Fig. 1 B) Dendrogram of clustered libraries. (Fig. 1 C) Multidimensional scaling indicating the relatedness of the nine libraries.
- [10] **Fig. 2- Fig. 2C show clustering and multidimensional scaling of the 115 genes differentially expressed ( $p < 0.001$ ) in 9 SAGE libraries.** (Fig. 2A) Cluster of the 115 genes (left panel) with 3 main clusters (right panels) consisting of genes overexpressed in squamous cell carcinoma (upper), overexpressed in adenocarcinoma (middle) and underexpressed in adenocarcinoma (lower panel), respectively. † Indicates that this tag corresponds to more than one gene of the same family. \* Indicates that this tag corresponds to more than one distinct gene. (Fig. 2B) Dendrogram of 9 clustered libraries, using 115 differentially expressed genes. (Fig. 2C) Multidimensional scaling of the libraries, using 115 differentially expressed genes.

- [11] **Fig. 3A- Fig. 3B show a comparison of genes under-expressed in adenocarcinoma using Affymetrix GeneChips™ and SAGE libraries. (Fig. 3 A) Histogram of normalized SAGE data shows the average relative expression levels of 7 genes that were underexpressed in adenocarcinoma (shown in the lower right panel in Fig. 2C). (Fig. 3 B) Histogram of GeneChip™ data shows the normalized average relative expression levels of the same genes as in Fig. 3A. When a GeneChip™ expression value was less than 1, it was set to 1 before normalization. Normalization was done in the same manner as for clustering analysis.**

#### **DETAILED DESCRIPTION OF THE INVENTION**

- [12] It is a discovery of the present inventors that certain molecular markers can be used to distinguish between the two most common forms of lung cancer: adenocarcinoma and squamous cell carcinoma. By assessing the expression levels of certain genes in a sample tumor tissue relative to normal, non-pathological lung tissue, one can make a determination of which of these types the cancer represents.
- [13] Expression of any gene which has been found to be up-regulated or down-regulated in one or more cancer types can be measured. According to one preferred embodiment, a lung tissue can be diagnosed, prognosed, or treatment determined by ascertaining an expression pattern of one or more cancer markers. Such markers include, but are not limited to glutathione peroxidase (GPX; NM\_002083), glutathione S-transferase M3 (GSTM3; NM\_000849), aldoketoreductase family 1, member B 10 (NM\_020299), peroxiredoxin 1 (PRDX1; NM\_002574), small proline-rich protein 3 (SPRR3; NM\_005416), and TNF receptor superfamily member 18 (TNFRSF18; NM004195). The amount of the gene product determined in a suspected cancer tissue is compared to the amount of the same gene product in a lung tissue sample which is non-pathological. An increased or decreased amount of the gene product in the lung cancer sample relative to the lung tissue sample which is non-pathological identifies the lung cancer by type. Using such markers, one can distinguish between squamous cell carcinoma and adenocarcinoma of the lung, for example.

- [14] Either mRNA or protein can be measured as a means of determining up- or down-regulation of a gene. Any technique known in the art for measuring such gene products can be used. Quantitative techniques are preferred, however semi-quantitative or qualitative techniques can also be used. Suitable techniques for measuring gene products include, but are not limited to SAGE analysis, DNA microarray analysis, Northern blot, Western blot, immunocytochemical analysis, and ELISA,
- [15] Control samples which can be used according to the present invention include any non-pathological sample of lung tissue. These can be isolated from the same individual as the suspected lung sample or from a different individual, whether related or not. Suitable cell types include lung small airway epithelial cells as well as bronchial/tracheal epithelial cells.

### **Examples**

#### **Example 1. Tumors and cell lines.**

- [16] Primary lung tumor tissues used for SAGE were obtained from Johns Hopkins Hospital following surgery for lung resection due to cancer, and as previously described (9). Histologically, the two squamous tumors were moderately differentiated squamous cell carcinomas while the two adenocarcinomas consisted of a well differentiated and a poorly differentiated tumor with a shared common feature of lymphoplasmacytic infiltrations in the adjacent alveolar septa. SAEC and NHBE cells were purchased from Clonetics/BioWhittaker, Inc. (Walkersville, MD) and propagated following the manufacturer's instruction. We chose these two primary cell cultures as normal controls because they represented pure populations of lung epithelial cells from the small and large airways, respectively. Tumor RNA samples were either purchased from BioChain Inc. (Hayward, CA) or obtained in the same manner as samples used for SAGE (9). A549 cells were obtained as a gift from Dr. James Herman (Johns Hopkins Oncology Center).

#### **Example 2. SAGE libraries and SAGE analysis.**

- [17] Total RNA samples were isolated by RNazol B (Tel-Test Inc., Friendswood, Texas) according to the manufacturer's recommendations. Poly (A)<sup>+</sup> RNA was extracted using

the Oligotex mRNA Mini Kit (Qiagen Inc., Valencia, California) and the Dynabeads mRNA DIRECT Kit (Dynal A.S., Oslo, Norway). SAGE libraries were generated and the tags sequenced as described (11) (22). SAGE 300 software (URL address: [http file type, www host server, domain name sagenet.org, directory sage\\_protocol, subdirectory htm](http://www.sagenet.org/sage_protocol.htm)), was used to identify tag sequences and to quantify the abundance of each tag. The gene identity and UniGene cluster assignment of each SAGE tag was obtained using the tag-to-gene 'reliable' map (updated April 23, 2001) from URL address: [http file type, www host server, domain name ncbi.nlm.nih.gov, directory pub, subdirectory SAGE, subsubdirectory map](http://www.ncbi.nlm.nih.gov/pub/SAGE/map) and the table of UniGene clusters (updated May 23, 2001), from URL address: [http file type, www host server, domain namencbi.nlm.nih.gov, directory UniGene](http://www.ncbi.nlm.nih.gov/UniGene).

**Example 3. Normalization and hierarchical clustering analysis.**

- [18] The "Cluster 2.11" program (URL address: [http file type, domain name rana.lbl.gov](http://rana.lbl.gov/Cluster2.11)), was used for normalization and clustering of the SAGE data. Briefly, the normalization included logarithmic transformation of the data, followed by 10 cycles of centering the data on the median by samples, then by genes, each time scaling the sum of the squares in each sample and each gene to 1. The non-centered Pearson correlation was used for distance calculations and the weighted-average linkage was used for clustering as described (21).

**Example 4. Multidimensional scaling of normal lung and tumor samples.**

- [19] A classical multidimensional scaling method was used to determine the relatedness of each library analyzed by SAGE (23). Each sample was used to generate a unique library. A table of normalized expression levels for each gene in every library was used as a dissimilarity matrix. Normalization was performed using the "Cluster 2.11" program, as described above. Multidimensional scaling allows for the calculation of coordinates of objects if the distances between objects are known. The distances between the samples were calculated as  $1 - C_{nm}$ , where  $C_{nm}$  was the correlation coefficient between libraries  $n$  and  $m$ . The distance matrix spans an N-dimensional space, where N is the number of

libraries in the study. Principal Component Analysis (23) was used to best fit the libraries into a 3-dimensional realm for presentation purposes.

**Example 5. Statistical analysis.**

- [20] The p-chance analysis [available in the SAGE 300 software and described in (21)] was used to select genes most differentially expressed between each tumor and its corresponding normal controls. P-chance uses the Monte-Carlo method (24) to calculate the relative probability of detecting an expression difference equal to, or greater than, the observed expression difference between two samples by chance alone. For each tumor type, one of the two tumor libraries was first compared with the two corresponding normal libraries to select genes with a p-chance value of  $<0.001$ . At this p-chance, the false positive rate for all selected genes was  $<0.015$ . We next selected only those genes with consistent expression patterns in both tumor libraries of the same cell type and combined them with genes selected from the other tumor type using the same method.

**Example 6. Real-time quantitative PCR analysis.**

- [21] Five genes identified by SAGE as highly expressed in either adenocarcinomas or squamous cell carcinoma were analyzed by Real-time reverse transcription (RT)-PCR using 14 RNA samples from lung tumors and controls (25). The Real-time RT-PCR probes and primers were designed using Primer Express software (PE Biosystems, Foster City, California). Primer sequences and reaction conditions are described in the supplemental material. The relative expression of each gene was calculated as the ratio of the average gene expression levels for tumors of the same cell type compared to its corresponding normal.

**Example 7. Gene expression analysis using GeneChip™.**

- [22] GeneChip™ U95A probe arrays were obtained from Affymetrix Inc. (Santa Clara, CA). A total of 32 RNA samples were individually prepared, hybridized to the GeneChip™, and scanned by a Hewlett-Packard (HP) GeneArray™ scanner as recommended by the manufacturer. Six internal GeneChip™ standards,  $\beta$ -actin, 18S rRNA, 28S rRNA,



glyceraldehyde-3-phosphate dehydrogenase, transferrin receptor, and the transcription factor ISGF-3, were used as controls to ensure the quality of all samples tested.

#### **Example 8. SAGE of NSCLC.**

- [23] A total of nine independent SAGE libraries were generated from five different normal and tumor tissues. A total of 18,300 independent clones were sequenced to generate 374,643 tags that represented 66,501 distinct transcripts (Table 1). Of the 23,056 distinct tags that appeared more than once in all nine libraries combined, 18,595 tags had at least one match to a UniGene cluster, 4,907 tags had multiple matches, 4,319 tags had no match, and 142 tags matched to mitochondrial DNA or ribosomal RNA sequences. Accounting for 7% potential sequencing errors (21) in tags that appeared only once in all nine libraries, the total number of distinct transcript tags identified is about 59,000. Although this number exceeds the current estimate of 30,000 to 40,000 genes predicted in the human genome(26, 27), the discrepancy could be accounted for by alternatively spliced transcripts and polyadenylation usage sites, which can result in multiple SAGE tags for the same gene (28, 29). Alternatively, since our transcript analysis was done on a limited number of tissues, it is possible that the current gene estimates are low.

**Table 1. SAGE in NSCLC and normal lung bronchial epithelial cells.**

<b>Tissue Source</b>	<b>No.Clones</b>	<b>No.Tags</b>
Normal Human Bronchial Epithelial Cells-1 (NHBE-1)	3759	58,273
Normal Human Bronchial Epithelial Cells-2 (NHBE-2)	4046	59,885
Normal Small Airway Epithelial Cells-1 (SAEC-1)	838	21,318
Normal Small Airway Epithelial Cells-2 (SAEC-2)	1299	26,956
Squamous Cell Carcinoma-A	2259	56,817
Squamous Cell Carcinoma-B	2186	51,901
Adenocarcinoma-A	799	21,714
Adenocarcinoma-B	928	24,018
Adenocarcinoma cell line A549	2186	53,752

<b>Total Number</b>	18,300	374,634
---------------------	--------	---------

**Summary:**

No. unique libraries = 9

No. unique tags = 66,502

No. of unique tags that appear >1 = 23,056

No. matched to unique UniGene cluster = 18,652

**Example 9. Hierarchical clustering of tumor and normal lung tissues based on SAGE.**

[24] To identify genes that are differentially expressed between the tumors and the normal samples, as well as between the different tumor types, we examined the overall similarities of the libraries derived from each tissue using hierarchical clustering (22). Since expression differences for more commonly expressed genes are less likely to have been observed by chance, a collection of 3,921 SAGE tags appearing at least 10 times in all nine libraries was subjected to the clustering analysis. Although each sample was derived from a different individual and had a unique expression pattern (Fig. 1A), the normal tissues were more similar to each other and the tumor tissues were more alike as a group. Furthermore, the SAEC and NHBE samples each paired together under the normal branch, while the adenocarcinomas and the squamous cell tumors each clustered together under the tumor branch (Fig. 1B). The adenocarcinoma-derived A549 cell line branched with the NSCLC tumors and demonstrated its relatedness to the two adenocarcinomas in multi-dimensional scaling (23), which displays the spatial relationship of all nine samples with respect to one another (Fig. 1C).

[25] Because gene expression levels were represented by a tag-count for each transcript detected in the SAGE libraries, we used the Monte-Carlo simulation (24) to quantify the significance of gene expression differences between the tumor libraries and the two corresponding normal epithelial cell controls. At a  $p < 0.001$ , fifty-eight genes were selected when comparing the two adenocarcinomas to the two SAEC samples, and 71 genes were obtained by comparison of the squamous cell carcinomas to the NHBE cells. Because 14 genes were common to both of comparisons, we therefore identified 115 highly differentially expressed transcripts for both tumor types (Table 1, List of genes in

*Supplemental Material*). As expected, when subjected to hierarchical clustering, these 115 genes again separated the nine libraries into the exact same branching patterns (Fig. 2A) as with the nearly 4,000 genes described above. Once again, the A549 cell line branched with the tumor tissues and was located closest to the two adenocarcinomas by multi-dimensional scaling (Fig. 2B).

**Example 10. Biologically distinct clusters of genes in different NSCLC subtypes.**

- [26] The clustering of the 115 statistically significant genes revealed at least three distinct gene clusters that were highly characteristic of the tumor tissues analyzed (Fig. 2C). Genes most highly expressed in squamous carcinomas of the lung (Fig. 2C, upper panel) were characterized by transcripts encoding proteins with detoxification and antioxidant properties. These genes include glutathione peroxidase 2 (GPX2), glutathione S-transferase M1 (GSTM1), carboxylesterase, aldo-keto reductase, and peroxiredoxin 1 (PRDX1). Their presence in squamous cell lung cancers most likely represented cellular response by the bronchial epithelium to environmental carcinogenic insults (30, 31). The clustering of these overexpressed genes highlight the notion that functional variation of these proteins in the population may contribute to lung cancer susceptibility in some patients. In addition, GSTM1 is a known susceptibility marker for lung and oral cavity cancer (32). It has also been associated with breast (33) and ovarian cancers (34). Interferon alpha-inducible protein 27 is also shown to be overexpressed in 50% of breast cancers (35).
- [27] In contrast, the cluster of genes overexpressed in lung adenocarcinoma (Fig 2C, middle panel) mostly encoded small airway-associated proteins and immunologically related proteins. The presence of surfactants A2 and B, pronapsin A, and mucin1 in the cluster reflects the origin of tumors derived from small airway epithelial cells, such as type 2 pneumocytes and Clara cells (36, 37). However, high expression of these genes also suggested that these proteins may participate in the tumorigenesis of lung adenocarcinomas. Indeed, mucin1 is also overexpressed in breast cancers and tyrosine phosphorylation of the CT domain of MUC1 mucin leads to activation of a mitogen-activated protein kinase pathway through the Ras-MEK-ERK2 pathway (38, 39). Furthermore, the overexpression of immunoglobulin genes in adenocarcinomas examined

may be explained by the extent of B-cell infiltration and the presence of antigen presenting cells (APC) in the adenocarcinomas used for SAGE analysis. However, clustering analyses of the SAGE tags revealed that different tumor types preferentially expressed a different set of cell surface markers. Squamous cell cancers appeared to overexpress MHC class I and CD71 proteins (Fig. 2C upper panel), while adenocarcinomas had a relatively high expression of MHC class II and CD 74 antigens. This gene expression differences in tumors indicated that immuno-based cancer therapy might be augmented based on the expression of different tumor surface markers.

- [28] No unexpected, many of the genes underexpressed in the primary adenocarcinomas and the A549 adenocarcinoma cell line (Fig. 2C, lower panel) were those that are associated with squamous differentiation. These proteins include S100 proteins, keratins, and the small proline-rich protein 1B (Cornifin). Interestingly, two p53-inducible genes, 14-3-3 $\sigma$  (Stratifin) (40) and p21<sup>waf1/CIP1</sup> (41, 42), clustered with this group of genes, showing significantly reduced expression in adenocarcinomas. Both p21<sup>waf1/CIP1</sup> and 14-3-3 $\sigma$  are highly induced in cells treated with ionizing radiation and other DNA damaging agents in a p53-dependent manner (43, 44). Induction of these genes by p53 leads to cell cycle arrest (45). The p53 gene is frequently mutated in squamous carcinomas of the lung, and it is thought that mutations in p53 may contribute to the inability of lung epithelial cells to repair carcinogen-induced damage (46). In contrast, p53 mutations are observed much less frequently in lung adenocarcinomas (5). The reduced expression of both p21<sup>waf1/CIP1</sup> and 14-3-3 $\sigma$  gene transcripts in adenocarcinomas suggests that inactivation of genes in the p53-pathway play an important role in this lung tumor type as well. However, reduced expression of an mRNA may not always correlate with a reduction of the gene product. Further studies correlating the molecular status of p53 with the expression of the encoded proteins are needed to assess the involvement of p53 and its downstream genes in the development of lung adenocarcinoma.

#### **Example 11. Other genes differentially expressed in NSCLC.**

- [29] It is important to note that the 115 highly differentially expressed genes we have identified only represented a set of genes whose differential expression could distinguish the molecular characteristics of each cell type as well as the neoplastic condition in the

lung. Clearly, additional genes with biological significance to NSCLC could also be identified depending on the statistical method and the level of significance chosen. For example, when all tags that showed consistent expression within the libraries of the same cell type were compared to identify genes differentially expressed with a 99% confidence level, a larger number of candidate genes were identified. Specifically, 827 tags showed statistically significant differential expression between the squamous cell carcinomas and the NHBEs, with 71 tags showing at least 10-fold overexpression. A similar comparison of the two adenocarcinoma tumor libraries and the SAECs identified 298 tags showing differential expression, with 20 tags overexpressed at least 10-fold in the tumors. Jointly, 45 tags were differentially expressed in both comparisons and these genes were either a part of or further extended the observations revealed by the 115 genes. For example, small proline rich protein 3 (SPRR3) was elevated in the squamous tumors but was virtually absent in the adenocarcinomas. SPRR3 is a member of the small proline rich family of proteins which includes SPRR1 (Cornifin), a gene previously identified as a marker for squamous cell carcinoma (47) and is within the same cluster for genes underexpressed in adenocarcinomas (Fig. 2C lower panel). SPRR3 is a member of the proteins in the cornified cell envelope that help provide a protective barrier to the epidermal layer of cells (48). Reduced expression of this family of proteins in adenocarcinoma may contribute to the invasive properties of this cancer. Moreover, several members of the tumor necrosis factor (TNF) family of proteins and their receptors have demonstrated increased expression in various cancers including NSCLC (49). Our statistical analysis of the SAGE data revealed that expression of the *TNF receptor superfamily member 18* gene was increased in squamous cell tumors in addition to the detoxification and antioxidation genes. TNF promotes T-cell mediated apoptosis (50) and elevated expression of genes in this pathway may provide a mechanism for anti-proliferation of the tumor cells.

**Example 12. Quantitative PCR and GeneChip<sup>TM</sup> cDNA oligoarray analyses of additional NSCLC tumors.**

- [30] Because SAGE libraries were derived from only selected tumor tissues, it was essential to determine whether gene expression patterns derived from SAGE could be reproduced

in larger panel lung tissues using independent assays. A total of 43 additional tumor and normal samples were examined using either quantitative real-time PCR or cDNA arrays methods. Five genes observed by SAGE as highly overexpressed in either squamous or adenocarcinomas of the lung (listed in Fig. 2C) were examined by Real-time RT-PCR using 10 different NSCLC tumors and four normal controls. As shown in Table 2, Real-time RT-PCR indicated that the two squamous-tumor specific genes had consistently high expression ratios in this tumor type compared to its expression in adenocarcinomas. Similarly, the three adenocarcinoma-specific genes had consistently higher expression in this tumor type and much less in squamous cell cancers compared to the normal.

Table 2. Real-time quantitative PCR analysis of SAGE-identified genes.

Spec.	Tag	Accession	Description	No. of SAGE tags in library*										Ave. RT PCR†	
				N1	N2	S1	S2	Sq A	Sq B	Ad A	Ad B	Sq/N	Ad/S		
Sq	<u>GGTGGTGTCT</u>	<u>X53463</u>	Glutathione peroxidase 2 (GPX2)	4	2	0	1	58	41	0	0	11	2		
Sq	GCCCCCTTCC	AF241229	TNF receptor superfamily member 18	0	1	0	0	11	8	0	0	38	5		
Ad	GAAATAAAGC	Y14737	Ig heavy constant gamma 3	0	0	0	0	5	1	293	23	1	17		
Ad	GTTACACATTA	AI248864	CD74 antigen	0	1	0	1	9	2	86	21	31	93		
Ad	GGGCATCTCT	J00196	Major histocompatibility complex, class II	0	0	0	0	1	1	51	19	275	1800		

Expression of the listed genes was examined in 14 samples, including five squamous cell tumors, four adenocarcinomas, one tumor with adenocarcinoma morphology, two NHBE cultures and two SAEC cultures. \* The actual number of tag occurrences in the indicated SAGE library is provided. †The average expression of each gene was calculated for the four distinct cell types, and the ratio of differential expression is indicated. Ad = Adenocarcinoma, Sq = Squamous cell carcinoma, N = NHBE, S = SAEC, Spec. = Tumor specificity based on SAGE.

- [31] In order to survey the overall reliability of the molecular clustering obtained from lung SAGE libraries, we used GeneChip<sup>TM</sup> cDNA oligoarray (15, 16) to survey 32 tumor and normal samples (including three samples used in Real-time PCR) for relative gene expression. Only 51 of 115 highly differentially expressed transcript tags were present in the 12,000 element GeneChip<sup>TM</sup> (U95A), and 20 of 35 genes from the three main clusters (shown in Fig. 2C) were comparable by both SAGE and the cDNA array. The gene expression levels for these 20 genes were averaged among all tumors of the same cell type and compared to that of the corresponding normal samples. Nineteen of 20 genes displayed an expression pattern similar to those obtained by SAGE. The expression patterns for the cluster of genes down-regulated in adenocarcinomas are shown (Fig. 3A and 3B). These results indicate that hierarchical clustering of the SAGE libraries can reveal gene clusters with strong biological significance and support the notion that the highly quantitative and reproducible nature of SAGE can result in highly precise tissue classification and reliable gene clustering, using only a few tissue samples. Furthermore, because SAGE method is independent of the knowledge of the gene sequence or the probe hybridization condition, it allows for an unbiased identification and quantification of gene expression patterns in the tissues of interest. The use of SAGE can offer the opportunity to identify novel genes and molecular markers
- [32] In summary, we have used SAGE and hierarchical clustering analysis to identify molecular profiles and clusters of genes specifically associated with two of the most common types of human lung cancer. Although biologically significant and highly reproducible, the gene expression profile described here may only represented the basic molecular features from which adenocarcinoma and squamous cell carcinoma of the lung can potentially be distinguished. Histological features and clinical behavior of the tumor may depend on less pronounced changes in expression levels for a variety of genes and pathways. Nevertheless, cumulating evidence suggests that gene expression patterns most likely determine the clinical behavior and therapeutic response of the cancer (19, 51). The list of highly differentially expressed genes that we described will likely provide new molecular targets for improved diagnosis, prognosis, and rational therapy.



The analyses for the expression of these in a larger number of lung tumors with detailed clinical information and outcome will be help to accomplish this goal.

- [33] While the invention has been described with respect to specific examples including presently preferred modes of carrying out the invention, those skilled in the art will appreciate that there are numerous variations and permutations of the above described techniques that fall within the spirit and scope of the invention as set forth in the appended claims.

## References

1. AmericanCancerSociety (2001) *Cancer Facts and Figures 2001*, Atlanta Georgia).
2. Travis, W. D., Linder, J. & Mackay, B. (1996) in *Lung Cancer Principles and Practice*, eds. Pass, H. I., Mitchell, J. B., Johnson, D. H. & Turrisi, A. T. (Lippincott-Raven, New York), pp. 361-395.
3. Gazdar, A. F. (1994) *Anticancer Res* **14**, 261-7.
4. Graziano, S. L., Gamble, G. P., Newman, N. B., Abbott, L. Z., Rooney, M., Mookherjee, S., Lamb, M. L., Kohman, L. J. & Poiesz, B. J. (1999) *J Clin Oncol* **17**, 668-75.
5. Niklinska, W., Chyczewski, L., Laudanski, J., Sawicki, B. & Niklinski, J. (2001) *Folia Histochem Cytobiol* **39**, 147-8.
6. Bennett, W. P., Hussain, S. P., Vahakangas, K. H., Khan, M. A., Shields, P. G. & Harris, C. C. (1999) *J Pathol* **187**, 8-18.
7. Hainaut, P. & Pfeifer, G. P. (2001) *Carcinogenesis* **22**, 367-74.
8. Forgacs, E., Zochbauer-Muller, S., Olah, E. & Minna, J. D. (2001) *Pathol Oncol Res* **7**, 6-13.
9. Hibi, K., Liu, Q., Beaudry, G. A., Madden, S. L., Westra, W. H., Wehage, S. L., Yang, S. C., Heitmiller, R. F., Bertelsen, A. H., Sidransky, D., *et al.* (1998) *Cancer Res* **58**, 5690-4.
10. Gray, J. W. & Collins, C. (2000) *Carcinogenesis* **21**, 443-52.
11. Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. (1995) *Science* **270**, 484-7.
12. Polyak, K., Xia, Y., Zweier, J. L., Kinzler, K. W. & Vogelstein, B. (1997) *Nature* **389**, 300-5.

13. He, T. C., Sparks, A. B., Rago, C., Hermeking, H., Zawel, L., da Costa, L. T., Morin, P. J., Vogelstein, B. & Kinzler, K. W. (1998) *Science* **281**, 1509-12.
14. Hermeking, H., Rago, C., Schuhmacher, M., Li, Q., Barrett, J. F., Obaya, A. J., O'Connell, B. C., Mateyak, M. K., Tam, W., Kohlhuber, F., *et al.* (2000) *Proc Natl Acad Sci U S A* **97**, 2229-34.
15. DeRisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A. & Trent, J. M. (1996) *Nat Genet* **14**, 457-60.
16. Jordan, B. R. (1998) *J Biochem (Tokyo)* **124**, 251-8.
17. Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., *et al.* (2000) *Nature* **403**, 503-11.
18. Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., *et al.* (2000) *Nature* **406**, 747-52.
19. Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O. P., *et al.* (2001) *N Engl J Med* **344**, 539-48.
20. Notterman, D. A., Alon, U., Sierk, A. J. & Levine, A. J. (2001) *Cancer Res* **61**, 3124-30.
21. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. (1998) *Proc Natl Acad Sci U S A* **95**, 14863-8.
22. Zhang, L., Zhou, W., Velculescu, V. E., Kern, S. E., Hruban, R. H., Hamilton, S. R., Vogelstein, B. & Kinzler, K. W. (1997) *Science* **276**, 1268-72.
23. Cox, T. F. & Cox, M. A. (2001) *Multidimensional Scaling* (Chapman and Hall / CRC, New York).
24. Hammersley, J. M. & Handscomb, D. C. (1964) *Monte Carlo Methods* (Wiley, New York).

25. Higuchi, R., Fockler, C., Dollinger, G. & Watson, R. (1993) *Biotechnology* **11**, 1026-1030.
26. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001) *Nature* **409**, 860-921.
27. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., *et al.* (2001) *Science* **291**, 1304-1351.
28. Mironov, A. A., Fickett, J. W. & Gelfand, M. S. (1999) *Genome Res* **9**, 1288-93.
29. Brett, D., Hanke, J., Lehmann, G., Haase, S., Delbruck, S., Krueger, S., Reich, J. & Borka, P. (2000) *FEBS Lett* **474**, 83-6.
30. Auerbach, O. (1980) in *Pulmonary Diseases and Disorders*, ed. Fishman, A. P. (McGraw-Hill, New York), pp. 1388-1396.
31. Sekido, Y., Fong, K. M. & Minna, J. D. (1998) *Biochim Biophys Acta* **1378**, F21-59.
32. Nair, U. & Bartsch, H. (2001) *IARC Sci Publ* **154**, 271-90.
33. Mitrunen, K., Jourenkova, N., Kataja, V., Eskelinen, M., Kosma, V. M., Benhamou, S., Vainio, H., Uusitupa, M. & Hirvonen, A. (2001) *Cancer Epidemiol Biomarkers Prev* **10**, 229-36.
34. Howells, R. E., Holland, T., Dhar, K. K., Redman, C. W., Hand, P., Hoban, P. R., Jones, P. W., Fryer, A. A. & Strange, R. C. (2001) *Int J Gynecol Cancer* **11**, 107-112.
35. Rasmussen, U. B., Wolf, C., Mattei, M. G., Chenard, M. P., Bellocq, J. P., Chambon, P., Rio, M. C. & Basset, P. (1993) *Cancer Res* **53**, 4096-101.
36. Colby, T. V., Koss, M. N. & Travis, W. D. (1995), eds. Rosai, J. & Sobin, L. H. (Armed Forces Institute of Pathology, Washington, D.C.), pp. 10.
37. Chuman, Y., Bergman, A., Ueno, T., Saito, S., Sakaguchi, K., Alaiya, A. A., Franzen, B., Bergman, T., Arnott, D., Auer, G., *et al.* (1999) *FEBS Lett* **462**, 129-34.

38. Taylor-Papadimitriou, J., Burchell, J., Miles, D. W. & Dalziel, M. (1999) *Biochim Biophys Acta* **1455**, 301-13.
39. Meerzaman, D., Shapiro, P. S. & Kim, K. C. (2001) *Am J Physiol Lung Cell Mol Physiol* **281**, L86-91.
40. Hermeking, H., Lengauer, C., Polyak, K., He, T. C., Zhang, L., Thiagalingam, S., Kinzler, K. W. & Vogelstein, B. (1997) *Mol Cell* **1**, 3-11.
41. el-Deiry, W. S., Harper, J. W., O'Connor, P. M., Velculescu, V. E., Canman, C. E., Jackman, J., Pietenpol, J. A., Burrell, M., Hill, D. E., Wang, Y., *et al.* (1994) *Cancer Res* **54**, 1169-74.
42. Harper, J. W., Adami, G. R., Wei, N., Keyomarsi, K. & Elledge, S. J. (1993) *Cell* **75**, 805-16.
43. Waldman, T., Lengauer, C., Kinzler, K. W. & Vogelstein, B. (1996) *Nature* **381**, 713-6.
44. Chan, T. A., Hermeking, H., Lengauer, C., Kinzler, K. W. & Vogelstein, B. (1999) *Nature* **401**, 616-20.
45. Taylor, W. R. & Stark, G. R. (2001) *Oncogene* **20**, 1803-15.
46. Therrien, J. P., Drouin, R., Baril, C. & Drobetsky, E. A. (1999) *Proc Natl Acad Sci U S A* **96**, 15038-43.
47. Hu, R., Wu, R., Deng, J. & Lau, D. (1998) *Lung Cancer* **20**, 25-30.
48. De Heller-Milev, M., Huber, M., Panizzon, R. & Hohl, D. (2000) *Br J Dermatol* **143**, 733-40.
49. Tran, T. A., Kallakury, B. V., Ambros, R. A. & Ross, J. S. (1998) *Cancer* **83**, 276-82.
50. Holtzman, M. J., Green, J. M., Jayaraman, S. & Arch, R. H. (2000) *Apoptosis* **5**, 459-71.
51. Scherf, U., Ross, D. T., Waltham, M., Smith, L. H., Lee, J. K., Tanabe, L., Kohn, K. W., Reinhold, W. C., Myers, T. G., Andrews, D. T., *et al.* (2000) *Nat Genet* **24**, 236-44.

## We Claim:

1. A method of identifying a lung cancer as squamous cell carcinoma comprising:  
determining an amount of a gene product of a gene in a lung cancer sample, said gene selected from the group consisting of : glutathione peroxidase (GPX;; NM\_002083), glutathione S-transferase M3 (GSTM3; NM\_000849), aldo-ketoreductase family 1, member B 10 (NM\_020299), peroxiredoxin 1 (PRDX1; NM\_002574), small proline-rich protein 3 (SPRR3; NM\_005416), and TNF receptor superfamily member 18 (TNFRSF18; NM004195);  
comparing the amount of the gene product to the amount determined in a lung tissue sample which is non-pathological, wherein an increased amount of the gene product in the lung cancer sample relative to the lung tissue sample which is non-pathological identifies the lung cancer as a squamous cell carcinoma.
2. The method of claim 1 further comprising the step of using the determined comparative gene product information to formulate a diagnosis.
3. The method of claim 1 further comprising the step of using the determined comparative gene product information to formulate a prognosis.
4. The method of claim 1 further comprising the step of using the determined comparative gene product information to formulate a treatment plan.
5. The method of claim 1 wherein the gene product of GPX is determined.
6. The method of claim 1 wherein the gene product of GSTM3 is determined.
7. The method of claim 1 wherein the gene product of aldo-ketoreductase is determined.
8. The method of claim 1 wherein the gene product of PRDX1 is determined.

9. The method of claim 1 wherein the gene product of SPRR3 is determined.
10. The method of claim 1 wherein the gene product of TNF receptor 18 is determined.
11. The method of claim 1 wherein the gene product is mRNA.
12. The method of claim 1 wherein the gene product is protein.
13. The method of claim 1 wherein the lung tissue sample which is non-pathological comprises normal lung small airway epithelial cells.
14. The method of claim 1 wherein the lung tissue sample which is non-pathological comprises normal bronchial/tracheal epithelial cells.
15. The method of claim 1 wherein the amount of the gene product is determined using a microarray.
16. The method of claim 15 wherein cRNA is hybridized to probes on the microarray to determine the amount of the gene product.
17. A method of identifying a lung cancer as adenocarcinoma, comprising:
  - determining an amount of a gene product of a small proline-rich protein 3 (SPRR3; NM\_005416) gene in a lung cancer sample;
  - comparing the amount of the gene product to the amount determined in a lung tissue sample which is non-pathological, wherein a decreased amount of the gene product in the lung cancer sample relative to the lung tissue sample which is non-pathological identifies the lung cancer as adenocarcinoma.
18. The method of claim 17 further comprising the step of using the determined comparative gene product information to formulate a diagnosis.

19. The method of claim 17 further comprising the step of using the determined comparative gene product information to formulate a prognosis.
20. The method of claim 17 further comprising the step of using the determined comparative gene product information to formulate a treatment plan.
21. The method of claim 17 wherein the gene product is mRNA.
22. The method of claim 17 wherein the gene product is protein.
23. The method of claim 17 wherein the lung tissue sample which is non-pathological comprises normal lung small airway epithelial cells.
24. The method of claim 17 wherein the lung tissue sample which is non-pathological comprises normal bronchial/tracheal epithelial cells.
25. The method of claim 17 wherein the amount of the gene product is determined using a microarray.
26. The method of claim 25 wherein cRNA is hybridized to probes on the microarray to determine the amount of the gene product.



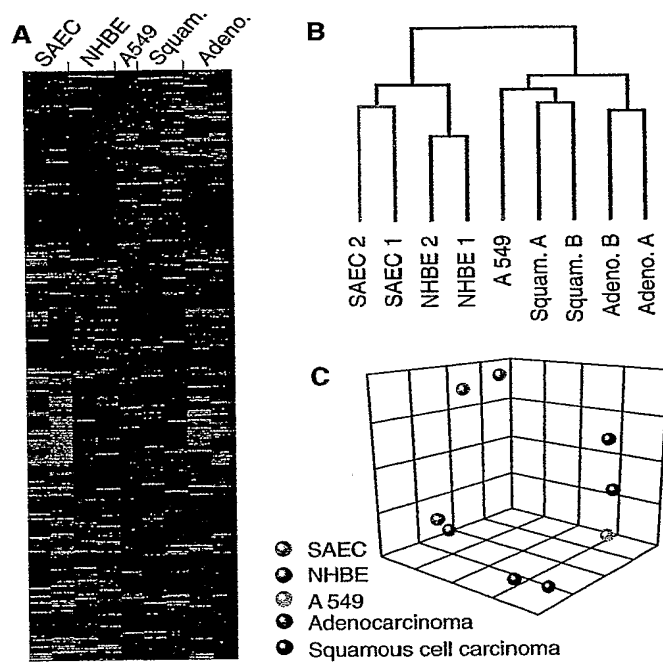


Fig. 1

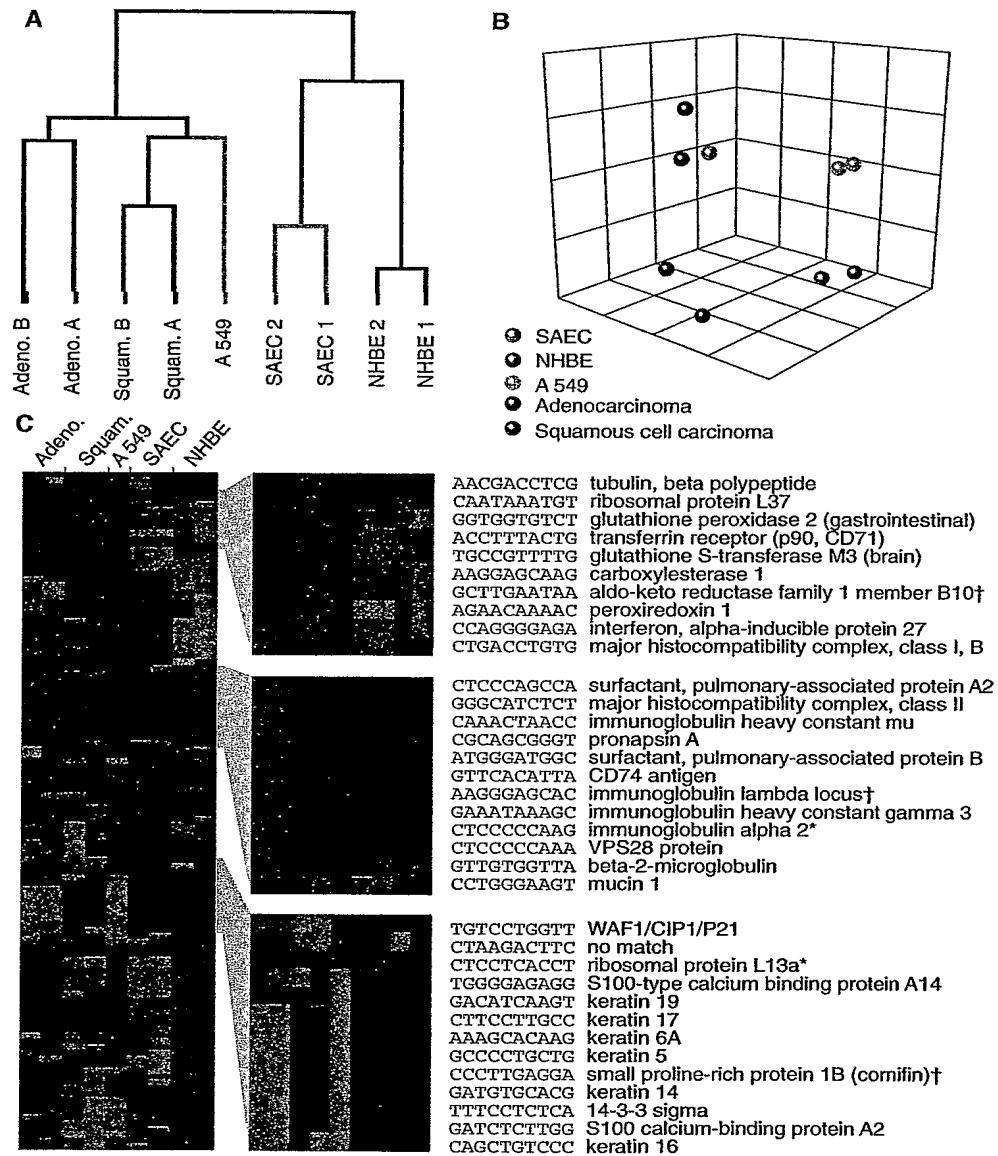


Fig. 2